**POPULATION HEALTH INFORMATION SYSTEM
RATES MACRO**

Manitoba Centre for Health Policy
and Evaluation

Version 6

TABLE OF CONTENTS

POPULIS RATES MACRO
(version 6)

1. INTRODUCTION

The Populis (Population Health Information System) Rates macro calculates rates of events per 1,000 population. Rates are calculated for user-defined subsets of the population (areas), as well as for the population as a whole. The macro generates crude, direct adjusted, and indirect adjusted rates. Statistics (standard errors, confidence intervals and a T^2 test of the null hypothesis that the rates are the same in all areas) may optionally be generated.

The (crude) rate of occurrence of an event is the number of events which occur in a well-defined population during a well-defined time interval, divided by the total amount of observation time accumulated during that time interval. Observation time is expressed in units such as person-years. At the Centre, the time interval of interest is usually one year, and the denominator of the rate is approximated by the number of individuals in the population at a single point in time.

The data sources that we use at MCHPE to determine the number of events typically only include those members of the population for which at least one event was observed, and so are not generally suitable to use as a source for the denominator. For example suppose we want to calculate the population rate of discharges from hospitals in Manitoba in 1991. The MCHPE hospital file for 1991 contains one observation (record) for each discharge that occurred in 1991. This file can be used to determine the numerator of the rate, the number of discharges. Individuals in Manitoba who did not go to a hospital during 1991 are not represented in the hospital file, so another source must be used for the denominator of the rate. For instance a registry file, containing one observation for each individual in the population at year end, 1991, may be used to calculate the denominator of the rate.

Rates are standardized (adjusted) so that 'fair' comparisons can be made between different populations or between different areas. The comparisons will be unaffected by difference in the adjusting factors. For instance it is well known that age is highly associated with the incidence rate of many diseases. If the crude rate of disease is higher in area A than area B, the difference might be due mainly to the fact that area A has an older population, rather than to other, more interesting, causes. We say that age is a confounding variable in this case. To calculate an adjusted rate, a standard population is required.   The adjusted rate for area A is the rate that would apply if the distribution of the confounding variable in area A were the same as in the standard population. If the rate for area B is adjusted in the same way, using the same standard population, 'valid' comparisons between the two regions can be made. The macro produces two common types of adjusted rates, direct and indirect.

2. <u>DATA SETS</u>

To use the rates macro, you must first construct two SAS data sets: a numerator and a denominator data set. The numerator data set must include at least one rate variable, whose value represents the number of events which occurred in the population during the time interval of interest. An area variable is required, one or more confounding variables may also be provided (usually age and/or gender). The denominator data set is used to determine the number of individuals in the population at some discrete point in time. It must also include area and confounding variables and these variables must be of the same type (character or numeric), the same length, and have the same name as in the numerator data set.

Optionally, the user may provide another data set: standard population or standard rates. These will be used to standardize the direct and indirect rates respectively. If either data set is omitted, the total denominator data set (ie all areas combined) will be used to standardize the corresponding rates.

a) <u>NUMERATOR DATA SET</u>

i) <u>Structure</u>

The numerator data set must be a SAS data set, and may be claims-based (possibly multiple records per individual), individual-based (one record per individual), or area/age/sex summarized (one record for each area/age/sex stratum of the population). If the data has been area/age/sex summarized, however, the standard error, confidence interval and T^2 statistic may not be valid (see section 3, stats=sum).

ii) <u>Variables</u>

The numerator data set must contain at least one rate variable, an area variable, and optionally one or more confounding variables. Depending on the structure of the data set, an individual variable may also be necessary. Optionally, a group variable may be included.

NOTE: Do not use a format statement to group values, as in the SAS statement `FORMAT age agefmt.;` Instead, create a new variable containing the formatted values, as in the statement `AGEGROUP=put(age,agefmt.));`

1. **RATE** variable(s), representing the number of events, for which rates are to be calculated. Several rates can be calculated simultaneously, but the user must ensure that the first five characters of the variable names are unique. Rate variables must be numeric and may contain any non-negative value. If the rate variable is binary (takes only two values), make sure that the variable has a value of 1 if the event has occurred, and 0 if it has not.

    e.g. the variable LOS (length of stay), representing number of days in hospital

    e.g. the variable HYST where '1' represents 'hysterectomy' and '0' represents 'no hysterectomy'

To produce rates of LOS for hysterectomy, you would need to do one of the following:
    a) include only hysterectomy events in the numerator data set and specify LOS as the rate variable when the macro is called.

    b) add a variable named HYSTLOS to the numerator data set which equals LOS when HYST=1, and 0 when HYST=0. Specify HYSTLOS as the rate variable.

    c) use the GROUP variable (see below) to separate hysterectomies and non-hysterectomies, generating rates for both values.

NOTE: If each observation in the numerator data set represents one event, the rate variable must take the value '1' for every observation. The macro will sum over this variable and so count the number of events. For example, suppose you are calculating the rate of discharges from hospital and each observation in the numerator data set represents one discharge. The numerator data set should include a variable named eg "DISCH" which equals '1' for every observation.

2. **AREA** variable, representing mutually exclusive subsets of the population, usually geographical sub-divisions. Note that no area should be contained in, or overlap with, another area. The variable may be numeric or character. The macro will calculate rates for each area separately, as well as rates for the total area (ie all areas combined). Optionally rates can be generated for a user-defined sub-total area (see section 4.12).

    e.g. the variable REGION, which takes values 'A' through 'J'

3. **CONFOUNDING** variable one, usually gender. The variable may be numeric or character.

    e.g. the variable SEX , where 'M' represents male, 'F' represents female

4. **CONFOUNDING** variable two, usually age, grouped into discrete units. The variable may be numeric or character.

    e.g. the variable AGEGROUP, where '1' represents ages 0-18, '2' represents ages 19-64, and '3' represents ages 65 and over.

5. **BYGROUP** variable (optional). Observations in the numerator data set will be grouped according to the values of this variable, and a rate will be calculated for each group separately. (This is the similar to using a BY statement in a SAS PROC step). The macro will also automatically calculate rates for the total group, that is over all values of the group variable.   The user should try to limit the number of levels of the GROUP variable to <100. This variable may be numeric or character.

    NOTE: If a bygroup variable is specified, and the numerator data set is summarized by stratum, it must have one observation for each area/sex/age/group stratum.

e.g., STAYTYPE - (type of hospital stay) where '1' = short-stay, '2' = long-stay, '3' = outpatient
The macro will calculate a separate rate for short-stay, long- stay and outpatient, and a total rate for all types of stay.

NOTE: If an individual has records that occur in more than one group, the combined rate may not be accurate.   Using the above example, suppose we are trying to calculate the rate of persons admitted to hospital.   Then if a single individual has both long and short stays, the total rate will be artificially inflated.

6. **INDIVIDUAL** variable, which takes a separate value for each individual in the data set. This variable is required if the data set includes multiple records per individual.

e.g. PHIN

b) <u>DENOMINATOR DATA SET</u>

i) <u>Structure</u>

The denominator data set must be a SAS data set. It can be a registry type data set with one record for each individual in the population, or it can be a summarized data set containing one record for each area/age/sex stratum and including a population count variable. If you are doing a lot of repetitive runs, a summarized denominator data set can dramatically reduce processing time.

Make sure that the population represented in the denominator data set is appropriate for the rate being calculated. If the numerator data set represents the number of events in a geographical region in 1992, then the denominator data set should represent the number of individuals in the same region at some point in 1992. If the numerator data set has exclusions (for example, only includes individuals over the age of 65) the denominator data set should exhibit similar restrictions. If you are calculating rates of hysterectomies, for example, you may want to report the rate per 1000 females, rather than per 1000 population.

ii) <u>Variables</u>

The denominator data set must include the same area and confounding variable(s) as the numerator data set. Depending on the structure of the data set, a population variable may also be necessary.

1. **AREA** variable, of the same type, length, and name as in the numerator data set.
   e.g. REGION

2. **CONFOUNDING** variable one, of the same type, length, and name as in the numerator data set.
   e.g. SEX

3. **CONFOUNDING** variable two, of the same type, length, and name as in the numerator data set.
e.g. AGEGROUP

4. **POPULATION** variable, representing the number of individuals in each area/sex/age stratum. If this variable is omitted when the macro is called, the macro assumes that the denominator data set is a registry file, with one record for each individual. In this case, the macro will create a variable named POP and use it to count the number of records in the data set.
   e.g. POPCOUNT

c) <u>STANDARD POPULATION DATA SET</u> (optional)

i) <u>Structure</u>

This data set is optional, and if supplied will be used to standardize the direct rates. It must be a SAS data set and it must have the same structure (registry or summarized) as the denominator data set.

ii) Variables

The standard population data set must include the same confounding variable(s) as the numerator data set. If the data set is summarized, it must contain a population variable with the same name as the population variable in the denominator data set.

1. **CONFOUNDING** variable one, of the same type, length, and name as in the numerator data set.
   e.g. SEX

2. **CONFOUNDING** variable two, of the same type, length, and name as in the numerator data set.
   e.g. AGEGROUP

3**. POPULATION** variable, representing the number of individuals in each stratum. It must have the same type, length and name as in the denominator data set. Omit this variable if the standard population data set is a registry file. In this case, the macro will create a variable named POP and use it to count the number of records in the data set.
   e.g. POPCOUNT

d) STANDARD RATES DATA SET (optional)

i) Structure

This data set is optional, and if supplied will be used to standardize the indirect rates. It must be a SAS data set and contain one observation for each confounding-bygroup combination.

ii) Variables

The standard rates data set must include the same confounding and bygroup variables as the numerator data set. There should be a variable that contains the standard rate per indvidual for each of the rate variable(s) in the numerator data set. These variables must have the same name as those in the numerator data set.

1. **RATE** variable(s). These variables must have the same name as those in the numerator data set, but contain the standard rate per individual.

2. **CONFOUNDING** variable one, of the same type, length, and name as in the numerator data set.
   e.g. SEX

3. **CONFOUNDING** variable two, of the same type, length, and name as in the numerator data set.
   e.g. AGEGROUP

4. **BYGROUP** variable (optional). If a group variable is defined for the numerator data set then then the standard rate data set must also contain the same variable.

3. <u>CALLING THE RATES MACRO</u>

The macro was written as a statment style macro (`options implmac`), it can be called by the
macro name followed by a series of options. As discussed below, the optons indicate names of
variables or data sets, values of variables, SAS expressions, or options.

To invoke the macro, use the following SAS statement:

```
pop_rate <options> ;
```

This statement must be placed after both numerator and denominator data sets have been defined. If an
option is not defined then the default will be used.

a) <u>OPTIONS</u>

Defaults are listed in the square backets following the option name.

var=[] - Rate variable name(s). Any number of rate variables are allowed and they do not have to
be mutually exclusive. If multiple variables are used they should be separated by spaces and
quoted. *Required*

```
var='HYST LOS CLAIM'
```

numdata=[] - Numerator data set name. *Required.*

```
numdata=HOSP92
```

popdata=[] - Denominator data set name. *Required*.

```
popdata=POP92
```

pop=[] - Name of population variable in the denominator data set. Only necessary if the
denominator data set is area/age/sex summarized.

```
pop=POPCOUNT
```

area=[region] - Area variable name.

```
area=REGION
```

conf1=[] - Name of confounding variable one, usually gender. This option must be quoted if
multiple variables are passed.

```
conf1=SEX
```

conf2=[] - Name of confounding variable two, usually age in discrete groups. This option must be quoted if multiple variables are passed.

```
conf2=AGEGROUP
```

bygrp=[] - Group variable name for numerator data.

```
bygrp=STAYTYPE
```

NOTE: If this option is not defined, the variable FGROUP, which takes only a single value (1), is created. This variable will appear in the printed output and output datasets. It is a programming artefact and can be safely ignored.

id=[] - Name of variable that identifies individuals.

```
id=PHIN
```

tgval=[Z] - Value for the group variable which will be used to identify the total bygrp. It must be a value greater than those being used by the bygrp variable. It can be numeric 0-99, or character 1-99 or A-Z, but must be the same length and type as the other values of the group variable.

```
tgval=9
```

taval=[Z] - Value for the area variable which will be used to identify the total area. It must be a value greater than those being used by the area variable. It can be numeric 0-99, or character 1-99 or A-Z, but must be the same length and type as the other values of the area variable.

```
taval=N
```

(REGION has values of 'A' to 'J', so the value 'N' can be used to represent the area total.)

compval=[] - Value for the area variable which will be used to idenitfy the sub-total area, defined in the next parameter. It must be a value that is greater than those being used by the area variable, but smaller than the total area. It can be numeric 0-99, or character 1-99 or A-Z, but must be the same length and type as the other values of the area variable.

```
compval=K
```

(Values of 'A' to 'J','N' have already been used for the variable REGION, so the value 'K' can be used to represent the sub-total area.)

compdef=[] - Expression representing the comparison area range. This expression specifies a range of values which will be used to define thecomparison area. If only a single value is specified, it will be interpreted as the upper limit of the range. If an expression is used it must be quoted.

```
compdef=" 'H' LE region LE 'J' "
compdef=C
compdef= " 'A' < region < 'F' "
```

NOTE: An '=' symbol can not be used, if necessary use the SAS mnemonic instead (i.e. EQ, NE, GE, LE).

print=[] - Print option. There are four different print options.
all - print all calculated variables.
direct, indirect, crude - print only the indicated rates, and associated statistics.

```
print=all
```

debug=[] - Debug option. *debug* - generates excessive output (SAS mprint and notes options), useful for debugging.

```
debug=debug
```

stats=[] - Statistics option. There are two different stats options.
stats - generates statistics (standard errors, confidence intervals, and T^2 tests). Use this option when the numerator data set has not been summarized by stratum.
sum - generates statistics. Use this option only when the numerator data set is area/age/sex/(/group) summarized AND the event in question is binary at the individual level, i.e. an individual may experience 0 or 1 events (e.g. hysterectomy, death).

```
stats=stats
```

NOTE: For non-binary events, valid statistics can only be produced if the numerator data set is not summarized, but rather claims-based or individual-based .

stpop=[] - Standard population data set name. This data set will be used to standardize the direct rates. If omitted, the total denominator data set will be used to standardize the direct rates.

```
stpop=STPOP
```

confid=[95] - confidence interval for each rate.

```
confid=99
```
  (yields 99% confidence intervals around each rate - an α of 0.01)

To do a Bonferronni adjustment for multiple comparisons use $(1-(\alpha/n))*100$. Where n is the number of areas.

```
confid=99.75
```
  (overall 95% confidence intervals with 20 regions and an α of 0.05)

asrates=[asrates] - Define the output data set name for the strata specific rates.

summdata=[summdata] - Define the output data set name for the summary data.

t2=[t2] - Define the output data set name for the T^2 statistics.

trans=[log] - Type of transformation for T^2 statistic. There are two option *log* and *nolog*.

```
trans=nolog
```

title=[] - You can define a single title for your output. If a title is not defined then the macro will generate its own titles based on the variables. The title must be quoted.

footnote=[] - You can define a single foot note for your output. If a footnote is not defined then the macro will generate its own footnotes based on the variables.

time=[1] - Define the number of time periods that the event data represents. For example rare events (death) are often collated over two years of data, to generate annual rates of deaths you would use the option `time=2.`

strate=[] - Standard rates data set name. This data set will be used to standardize the indirect rates. If omitted, the total denominator data set will be used to standardize the indirect rates.

```
strate=STRATES
```

b) <u>EXAMPLE</u>

The following example call will generate Total Days/1000 individuals in the population. A comparison region with the value of N has been defined. The directly standardized rates are standardized to the population in the SUMPOP92 data set.

```
pop_rate var=los
   numdata=hosp89
   popdata=sumpop
   pop=popcount
   stpop=sumpop92
   id=phin
   bygrp=staytype
   conf1='agegroup sex'
   area=region
   tgval=9
   taval=Q
   compval=N
   compdef="'A' <= region <= 'G'"
   print=all
   stats=stats
```

```
confid=99 ;
```

4. <u>OUTPUT</u>

a) <u>TABLES</u>

When the print option is specified, the macro generates tables in the listing file. A single table is constructed for each value of the group variable. The values of the area variable form the rows of each table. The following quantities are produced for each rate variable that was specified when the macro was called.   (In what follows, substitute the first five characters of the rate variable name for the string "xxxxx")

Variable	Description
yyyyyyyy (usually POP)	Population counts
O_xxxxx	Observed frequency of event
E_xxxxx	Expected number of events under the null hypothesis that the rate is the same in all regions.
ER_xxxxx	Expected rate per 1000 population of events under the null hypothesis that the rate is the same in all regions.
S_xxxxx	Standard morbidity ratio (observed frequency divided by expected)
C_xxxxx	Crude rate of event per 1000 population
D_xxxxx	Direct adjusted rate of event per 1000 population
I_xxxxx	Indirect adjusted rate of event per 1000 population

If the "stats" or "sum" option was specified when the macro was called, the following quantities will also be produced:

Type of Rate	Variable	Description
Crude rate	SC_xxxxx	Standard error
	LC_xxxxx	Lower limit for confidence interval
	UC_xxxxx	Upper limit for confidecne interval
	TC_xxxxx	T^2 statistic
	PC_xxxxx	Probability for T^2 statistic
Direct rate	SD_xxxxx	Standard error
	LD_xxxxx	Lower limit for confidence interval
	UD_xxxxx	Upper limit for confidecne interval
	TD_xxxxx	T^2 statistic
	PD_xxxxx	Probability for T^2 statistic
Indirect rate	SI_xxxxx	Standard error
	LI_xxxxx	Lower limit for confidence interval
	UI_xxxxx	Upper limit for confidecne interval
	TI_xxxxx	T^2 statistic
	PI_xxxxx	Probability for T^2 statistic

NOTES:

i)   The reported standard error corresponds to the rate per 1000 population.

ii) The confidence intervals are constructed using a log transformation, and are for rates per 1000 population.

iii) The T^2 statistic can be used to test the null hypothesis that the rate of the event is the same in each area, and is constructed using a log transformation.

iv) If we repeatedly generate data under the conditions of the null hypothesis, and calculate the T^2 statistic each time, then PROBABILITY represents the proportion of times that these T^2 statistics would be more extreme in absolute value than the T^2 statistic that was actually observed. A very small value of PROBABILITY is taken as evidence against the null hypothesis.

v)   For more information, on the log transformation and T^2 test, see the paper by Carriere and Roos.

b) OUTPUT DATA SETS

The macro will also create three temporary SAS data sets:

1. **SUMMDATA**

All variables produced in the printed tables are included in the SUMMDATA data set. The data set contains one observation for each possible combination of the levels of the group and region variables.

Variables - For each rate variable, the following quantities are calculated and stored in the SUMMDATA data set: observed number of events, expected number of events, standard morbidity ratio, crude rate, direct rate, indirect rate, expected rate. Reported rates are per 1000 individuals. If the SUM or STATS option is requested when the macro is called, then standard errors, confidence limits, T^2 statistics and corresponding probabilities will be produced for each type of rate (crude, direct, and indirect). The data set also includes group, region, and population variables.

Variable names - The variable names and values are identical to those produced in the printed tables.

2. **T2**

The T2 data set is created only if the SUM or STATS option is requested when the macro is called. It contains T^2 statistics and corresponding probabilities for testing the hypothesis that the rates are the same in each area. The data set contains one observation for each value of the group variable (including the group total).

Variables - For each rate variable, the data set includes a T^2 statistic and corresponding probability for the three types of rate (crude, direct and indirect).

Variable names - The variable names and values are identical to those produced in the printed tables.

3. **ASRATES**

The confounding variables effectively divide the population into strata, and the rate of occurrence of the event in each stratum is referred to as the specific rate. Note that the specific rates are not included in the printed tables. The data set contains one observation for each possible combination of the levels of the group, region, and confounding variables.

Variables - The ASRATES data set contains the observed number of events and specific rate for each rate variable. The specific rate is per individual. If the SUM or STATS option is requested, the standard error of the specific rate will also be included in the data set. Additional variables which appear in this data set are group, region, confounding variables, and population.

Variable names - In what follows, subsititute the first five characters of the rate variable name for the string "xxxxx"

Variable	Description
xxxxx	observed number of events
AS_xxxxx	specific rate per individual
SR_xxxxx	standard error of the rate

5. EXAMPLE of RATES

The following is an example of a SAS program that reads in two types of data sets and generates rates output by using the RATES macro. The output produced by this program follows.

```
**********************************************************************;
* STEP 1:  READ IN MACRO AND CREATE FORMATS                       *;
* Create formats both for a) grouping and b) labelling variables *;
* in both numerator and denominator.                             *;
**********************************************************************;

** The rates macro is now in the system library and does not need
   to be included ;
%include '/home/burchil/maclib/pop_rate';    *Read in the RATES macro ;

proc format;
value agegrp  0-59 = '1'                   /*The format agegrp will be used to group age*/
              60-74 = '2'
              75-79 = '3'
              80-84 = '4'
              85-89 = '5'
            90-high = '6';

value agelbl     1 = ' 0-59'               /* agelbl will be used to label agegroup*/
                 2 = '60-74'
                 3 = '75-79'
                 4 = '80-84'
                 5 = '85-89'
                 6 = '90-high';

value $region    'A' = 'A.Central   '/* $region will be used to label region*/
                 'B' = 'B.Eastman   '/* Note that 'Z' is TOTAL region */
                 'C' = 'C.Interlake '
                 'D' = 'D.Norman    '
                 'E' = 'E.Parkland  '
                 'F' = 'F.Thompson  '
                 'G' = 'G.Westman   '
                 'H' = 'H.Winnipeg W'
                 'I' = 'I.Winnipeg N'
                 'J' = 'J.Winnipeg S'
                 'Z' = 'Z.Manitoba  ';

value $hspcode '1' = 'Teaching/Urban'/* $hspcode will be used to label hospcode */
                 '0' = 'All other'          /* Note that 'Z' is TOTAL hospitals */
                 'Z' = 'All hospitals';
run;


**********************************************************************;
* STEP 2:  DENOMINATOR - POPULATION CODE                          *;
* The following program reads in a permanent population data set *;
* and creates a temporary population data set.                   *;
* It also creates a permanent population summary file by region, *;
```

```
* age, and sex.  This summary file can then be used repeatedly for  *;
* rates runs requiring the same age/sex/region breakdowns.           *;
* This entire step can be repeated for a second population data set  *;
* if standardization to a different population is desired.        *;
*********************************************************************;


libname popdir '/cpe/db';                  *Directory containing population dataset;
libname popnew '/home/werbenk/practice'; *Directory assigned for optional permanent
                                    permanent dataset (if saved);
data pop89;                               *Create a temporary population data set;
  set popdir.pop89;                       *Read in population data;
  length agegrp 3;                        *Set length of new age variable to 3;
  region=put(muncode,$reg10f.);          *Create area variable;
  agegrp=input(put(age,agegrp.),2.);     *Create age variable;
  retain popcnt 1;                        *Create population count variable;

  format region $region. agegrp agelbl.; * Attach labels to new variables;

run;


***************************************************************;
* OPTIONAL:                                             *;
* Save a permanent dataset of summarized population data.  *;
* The PROC SUMMARY statement counts all individuals       *;
* (frequency) within each of the following classes:        *;
* REGION, AGEGRP, and SEX.                               *;
***************************************************************;

proc summary nway;
 class region agegrp sex;
 var popcnt;
 output out=popnew.pop_sum (drop=_type_ _freq_) sum=;
run;


*********************************************************************;
* STEP 3:  NUMERATOR - EVENT CODE                              *;
* The following program reads in a permanent data set of hospital   *;
* claims, or events, and creates a temporary event file           *;
* containing area (REGION) and age (AGEGRP) variables that are     *;
* identical to those in the denominator.  A PROC CONTENTS on both   *;
* numerator and denominator files has shown SEX to be identical     *;
* as well.                                                       *;
*********************************************************************;


libname hosp '/dsd1/ruta';             *Directory containing event dataset;

data event;                              *Create a temporary event data set;
 set hosp.arrayeg;                      *Read in event data ;
  length agegrp 3;                       *Set agegrp to same length as for pop. data set;

  region=put(muncode,$reg10f.);          *Create area variable;
  agegrp=input(put(ageadm,agegrp.),1.);  *Create age variable;

  if hosp<=30 then hospcode='1';         *Create grouping variable of interest ;
  else hospcode = '0';
```

```
  if hospcode='1';                              *Optional: Save processing time if not
                                                 interested in hospcode=0;
  format region $region. agegrp agelbl. hospcode $hspcode. ;
 run;


*****************************************;
* STEP 4: INVOKE RATES MACRO  *;
*****************************************;

pop_rate var=los numdata=event popdata=popnew.pop_sum
        pop=pop conf1='sex agegrp' bygrp=hospcode
        print=all stats=stats ;
run;
```

RATES EXAMPLE OUTPUT:

1. Regional summary rates

A separate table is produced for each value of the group variable (HOSPCODE). In this example, there is only a single rate variable, LOS. Since the STATS option was specified when the macro was called, the printed tables include statistics. This information is saved int he temporary data set SUMMDATA.

```
Regional summary data (rates are per 1,000)

                                Variables are: los

------------------------------------- HOSPCODE=Teaching/Urban -------------------------------------------

REGION        POPCNT O_LOS    E_LOS  S_LOS C_LOS SC_LOS LC_LOS UC_LOS I_LOS SI_LOS LI_LOS UI_LOS D_LOS SD_LOS LD_LOS UD_LOS

A.Central      480320  2989  6330.31  0.47  6.22  1.40    4.00   9.69  5.89  1.33   3.78   9.17  5.87  1.27   3.84   8.98
B.Eastman      435760  2961  4879.85  0.61  6.80  0.95    5.16   8.94  7.57  1.06   5.75   9.96  7.30  1.05   5.50   9.69
C.Interlake    376780  2389  4621.03  0.52  6.34  0.67    5.15   7.81  6.45  0.68   5.24   7.94  6.38  0.68   5.18   7.87
D.Norman       125915   431  1249.83  0.34  3.42  0.87    2.08   5.63  4.30  1.09   2.62   7.08  4.39  1.38   2.37   8.15
E.Parkland     240740   876  3692.84  0.24  3.64  0.55    2.71   4.89  2.96  0.44   2.20   3.97  3.45  0.55   2.53   4.72
F.Thompson     241020  1547  1814.87  0.85  6.42  1.00    4.73   8.72 10.63  1.66   7.83  14.44  8.73  1.61   6.09  12.53
G.Westman      593870  4118  8864.01  0.46  6.93  0.68    5.72   8.41  5.79  0.57   4.78   7.03  6.00  0.55   5.01   7.18
H.Winnipeg W   744320 15054  9913.35  1.52 20.23  2.54   15.82  25.86 18.94  2.38  14.81  24.22 18.90  2.28  14.91  23.95
I.Winnipeg N 1183785 22025 14227.08  1.55 18.61  3.04   13.51  25.63 19.31  3.16  14.02  26.60 19.30  3.08  14.12  26.39
J.Winnipeg S 1257845 18463 15259.83  1.21 14.68  1.33   12.30  17.52 15.09  1.36  12.64  18.02 15.29  1.45  12.70  18.41
Z.Manitoba   5680355 70853 70853.00  1.00 12.47  0.79   11.01  14.13 12.47  0.79  11.01  14.13 12.47  0.79  11.01  14.13
```

```
--------------------------------------- HOSPCODE=All hospitals -------------------------------------

REGION        POPCNT O_LOS    E_LOS  S_LOS C_LOS SC_LOS LC_LOS UC_LOS I_LOS SI_LOS LI_LOS UI_LOS D_LOS SD_LOS LD_LOS UD_LOS

A.Central      480320  2989  6330.31  0.47  6.22  1.40    4.00   9.69  5.89  1.33   3.78   9.17  5.87  1.27   3.84   8.98
B.Eastman      435760  2961  4879.85  0.61  6.80  0.95    5.16   8.94  7.57  1.06   5.75   9.96  7.30  1.05   5.50   9.69
C.Interlake    376780  2389  4621.03  0.52  6.34  0.67    5.15   7.81  6.45  0.68   5.24   7.94  6.38  0.68   5.18   7.87
D.Norman       125915   431  1249.83  0.34  3.42  0.87    2.08   5.63  4.30  1.09   2.62   7.08  4.39  1.38   2.37   8.15
E.Parkland     240740   876  3692.84  0.24  3.64  0.55    2.71   4.89  2.96  0.44   2.20   3.97  3.45  0.55   2.53   4.72
F.Thompson     241020  1547  1814.87  0.85  6.42  1.00    4.73   8.72 10.63  1.66   7.83  14.44  8.73  1.61   6.09  12.53
G.Westman      593870  4118  8864.01  0.46  6.93  0.68    5.72   8.41  5.79  0.57   4.78   7.03  6.00  0.55   5.01   7.18
H.WinnipegW    744320 15054  9913.35  1.52 20.23  2.54   15.82  25.86 18.94  2.38  14.81  24.22 18.90  2.28  14.91  23.95
I.WinnipegN  1183785 22025 14227.08  1.55 18.61  3.04   13.51  25.63 19.31  3.16  14.02  26.60 19.30  3.08  14.12  26.39
J.WinnipegS  1257845 18463 15259.83  1.21 14.68  1.33   12.30  17.52 15.09  1.36  12.64  18.02 15.29  1.45  12.70  18.41
Z.Manitoba   5680355 70853 70853.00  1.00 12.47  0.79   11.01  14.13 12.47  0.79  11.01  14.13 12.47  0.79  11.01  14.13


                Variable   Explanation    Variable  Explanation    Variable  Explanation
                  o_xxx      observed # s    E_xxx    Expected # s   SC_xxx    Crude S.ERR
                  C_x        Crude rates     D_xxx    Direct Rates   SD_xxx    Direct S.ERR
                  I_xxx      Indirect Rates S_xxx     SMR            SI_xxx    Indirect S.ERR
```

2. Significance Tables

For each value of the group variable (HOSPCODE), the T^2 statistic used to test the null hypothesis that the crude rate of LOS is the same in all areas, is calculated. The T^2 statistic and associated probability for this test appear in the significance table, labeled TC_LOS and PC_LOS respectively. Here the reported probability is effectively zero, giving very strong evidence that at least one of the areas has a crude rate that is different from the other areas. Similar statistics are calcuated for the direct and indirect rates. This information is saved in the temporary data set T2

```
-------------------- HOSPCODE=Teaching/Urban --------------------

OBS     TI_LOS     PI_LOS      TD_LOS     PD_LOS      TC_LOS     PC_LOS

 1      256.036         0      232.729         0      239.829         0

------------------- HOSPCODE=All hospitals ---------------------

OBS     TI_LOS     PI_LOS      TD_LOS     PD_LOS      TC_LOS     PC_LOS

 2      256.036         0      232.729         0      239.829         0
```

3. NOTE on ASRATES data set

No tables are produced for strata-specific rates; however, a third temporary data set - "ASRATES" - is created that contains an AS_xxx variable: strata-specific rates for the xxx variable by each value of the area variable (in this case, age- and sex-specific rates of LOS for each value of REGION). The SR_xxx variable is the standard error of the strata specific rate.

The following PROC PRINT illustrates the first few observations in the **ASRATES** data set:

```
OBS     HOSPCODE        REGION          AGEGRP    SEX    POPCNT     LOS     SR_LOS     AS_LOS

 1    Teaching/Urban    A.Central       0-59       1     40208      727      48.09     0.01808
 2    Teaching/Urban    B.Eastman       0-59       1     38622      834      65.47     0.02159
 3    Teaching/Urban    C.Interlake     0-59       1     32317      495      10.15     0.01532
 4    Teaching/Urban    D.Norman        0-59       1     11615      137       1.26     0.01180
 5    Teaching/Urban    Parkland        0-59       1     19020      219       4.41     0.01151
 6    Teaching/Urban    F.Thompson      0-59       1     23863      677      44.58     0.02837
 7    Teaching/Urban    G.Westman       0-59       1     46972      725      16.19     0.01543
 8    Teaching/Urban    H.Winnipeg W    0-59       1     62293     2869     299.72     0.04606
 9    Teaching/Urban    I.Winnipeg N    0-59       1     99008     3739    1368.89     0.03776
10    Teaching/Urban    J.Winnipeg S    0-59       1    104131     2325     129.31     0.02233
11    Teaching/Urban    Z.Manitoba      0-59       1    478049    12747    1988.16     0.02666
12    Teaching/Urban    A.Central       0-59       2     38552      361       3.48     0.00936
13    Teaching/Urban    B.Eastman       0-59       2     36409      917      30.69     0.02519
14    Teaching/Urban    C.Interlake     0-59       2     30416      629       5.47     0.02068
15    Teaching/Urban    D.Norman        0-59       2     10865       57       0.35     0.00525
16    Teaching/Urban    Parkland        0-59       2     17809      336       8.24     0.01887
17    Teaching/Urban    F.Thompson      0-59       2     22263      651       7.79     0.02924
18    Teaching/Urban    G.Westman       0-59       2     45631      766       5.35     0.01679
19    Teaching/Urban    H.Winnipeg W    0-59       2     58332     2264     214.71     0.03881
20    Teaching/Urban    I.Winnipeg N    0-59       2     98632     6529    7929.63     0.06620
```

6. FORMULAS

n_{ia} = number of individuals in stratum [†] i, area a, denominator population

$n_i = \sum\limits_{a} n_{ia}$ = number of individuals in stratum i, denominator population

n_{is} = number of individuals in stratum i, standard[††] population

x_{jia} = observed number of events for individual j, in stratum i, area a

$x_{ia} = \sum\limits_{j=1}^{n_{ia}} x_{jia}$ = observed number of events in stratum i, area a

$x_i = \sum\limits_{a} x_{ia}$ = observed number of events in stratum i

$x = \sum\limits_{i} \sum\limits_{a} x_{ia}$ = total number of observed events

$n = \sum\limits_{i} \sum\limits_{a} n_{ia}$ = total denominator population

$$v_{ia} = n_{ia} \frac{\sum\limits_{j=1}^{n_{ia}} \left(x_{jia} - \frac{x_{ia}}{n_{ia}} \right)^2}{n_{ia} - 1} = \text{estimated variance of} \quad x_{ia}$$

$$\text{if } x_{ia} = 0 \text{ then } v_{ia} = n_{ia} \frac{\left(1 - \frac{1}{n_{ia}} \right)^2}{n_{ia} - 1}$$

$X^2{}_v$ = Chisquare random variable with v degrees of freedom

A = number of areas

$\Pr(B)$ = Probability of event B

z_a = the $(1-a) \times 100\%$ quantile of the standard normal distribution.

For a standard normal random variable Z, $\Pr(Z > z_a) = a$.

Specific rate in stratum i, area a

$$\frac{x_{ia}}{n_{ia}}$$

Standard error of specific rate in stratum i, area a

[†] The population is divided into strata by the confounding variables.
[††] If a standard population data set is not specified when the macro is called the total denominator population (combined over all areas) is used, i.e. $n_{is} = n_i$

$$\frac{\sqrt{v_{ia}}}{n_{ia}}$$

Observed number of events in area a

$$\sum_i x_{ia}$$

Expected number of events in area a

$$e_a = \sum_i \left(\frac{x_i}{n_i} n_{ia} \right)$$

if $e_a = 0$ then $e_a = 1$

Standard morbidity ratio in area a

$$\frac{\sum_i x_{ia}}{e_a}$$

Expected rate in area a

$$\frac{e_a}{\sum_i n_{ia}}$$

Crude rate in area a

$$C_a = \frac{\sum_i x_{ia}}{\sum_i n_{ia}}$$

Standard error of crude rate in area a

$$SC_a = \frac{\sqrt{\sum_i v_{ia}}}{\sum_i n_{ia}}$$

Direct rate in area a

$$D_a = \sum_i \left(\frac{x_{ia}}{n_{ia}} \frac{n_{is}}{\sum_i n_{is}} \right)$$

Standard error of direct rate in area a

$$SD_a = \sqrt{\sum_i \left(\frac{v_{ia}}{n_{ia}^2} \frac{n_{is}^2}{\left(\sum_i n_{is}\right)^2} \right)}$$

Indirect rate in area a

$$I_a = \frac{x}{n} \frac{\sum_i x_{ia}}{e_a}$$

Standard error of indirect rate in area a

$$SI_a = \frac{x}{n} \frac{\sqrt{\sum_i v_{ia}}}{e_a}$$

Upper $(1-\alpha)\times100\%$ Confidence interval for crude Rate in area a

$$C_a \exp\left(z_{a/2} \frac{SC_a}{C_a} \right)$$

Lower $(1-\alpha)\times100\%$ Confidence interval for crude Rate in area a

$$C_a \exp\left(-z_{a/2} \frac{SC_a}{C_a} \right)$$

Upper $(1-\alpha)\times100\%$ Confidence interval for direct Rate in area a

$$D_a \exp\left(z_{a/2} \frac{SD_a}{D_a} \right)$$

Lower $(1-\alpha)\times100\%$ Confidence interval for direct Rate in area a

$$D_a \exp\left(-z_{a/2} \frac{SD_a}{D_a} \right)$$

Upper $(1-\alpha)\times100\%$ Confidence interval for indirect Rate in area a

$$I_a \exp\left(z_{a/2} \frac{SI_a}{I_a} \right)$$

Lower $(1-\alpha)\times100\%$ Confidence interval for indirect Rate in area a

$$I_a \exp\left(-z_{a/2} \frac{SI_a}{I_a} \right)$$

If the T^2 statistic is log transformed

T^2 statistic for crude rate

$$TC = \sum_a \frac{\left(\ln(C_a) - \ln\left(\frac{x}{n}\right)\right)^2}{\frac{SC_a^2}{C_a^2}}$$

T^2 statistic for direct rate

$$TD = \sum_a \frac{\left(\ln(D_a) - \ln\left(\sum_i \frac{x_i}{n_i} \frac{n_{is}}{\sum_i n_{is}}\right)\right)^2}{\frac{SD_a^2}{D_a^2}}$$

T^2 statistic for indirect rate

$$TI = \sum_a \frac{\left(\ln(I_a) - \ln\left(\frac{x}{n}\right)\right)^2}{\frac{SI_a^2}{I_a^2}}$$

If T^2 statistic is not transformed

T^2 statistic for crude rate

$$TC = \sum_a \frac{\left(C_a - \left(\frac{x}{n}\right)\right)^2}{SC_a^2}$$

T^2 statistic for direct rate

$$TD = \sum_a \frac{\left(D_a - \left(\sum_i \frac{x_i}{n_i} \frac{n_{is}}{\sum_i n_{is}}\right)\right)^2}{SD_a^2}$$

T^2 statistic for indirect rate

$$TI = \sum_a \frac{\left(I_a - \left(\frac{x}{n}\right)\right)^2}{SI_a^{\ 2}}$$

Probability associated with T^2 statistic for crude rate .

$$PC = \Pr\left(X_{A-1}^2 \geq TC\right)$$

Probability associated with T^2 statistic for direct rate .

$$PC = \Pr\left(X_{A-1}^2 \geq TD\right)$$

Probability associated with T^2 statistic for indirect rate .

$$PC = \Pr\left(X_{A-1}^2 \geq TI\right)$$

References:

Breslow and Day, Statistical Mehtods in Cancer Research, International Agency for Research on Cancer, 1980.

Khan, Harold A., and Sempos, Christopher T.: Statistical Methods in Epidemiology. New York: Oxford University Press, 1989 p85-98.

Carriere, Keumhee, and Roos, Leslie L.: Comparing standardized rates of events. Am J Epidemiol 1994; 140:5 pp 472-482

Carriere, Keumhee, and Roos, Leslie L.: A Method of Comparison of Standardized Rates of Low-Incidence Events. Medical Care 1997; 35:1 pp 57-69

Acknowledgements: